Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors

Faiyaz Al Zamal and Wendy Liu and Derek Ruths
School of Computer Science
McGill University
Montreal, Quebec Canada

Abstract

In this paper, we extend existing work on latent attribute inference by leveraging the principle of homophily: we evaluate the inference accuracy gained by augmenting the user features with features derived from the Twitter profiles and postings of her friends. We consider three attributes which have varying degrees of assortativity: gender, age, and political affiliation.

Our approach yields a significant and robust increase in accuracy for both age and political affiliation, indicating that our approach boosts performance for attributes with moderate to high assortativity. Furthermore, different neighborhood subsets yielded optimal performance for different attributes, suggesting that different subsamples of the user's neighborhood characterize different aspects of the user herself. Finally, inferences using only the features of a user's neighbors outperformed those based on the user's features alone. This suggests that the neighborhood context alone carries substantial information about the user.

Introduction

Latent attribute inference methods use available, unstructured online data generated by individuals to infer demographic attributes such as age, ethnicity, and political orientation for individual and groups of users (e.g., (Pennacchiotti and Popescu 2011)).

In this paper, we use the signal present in a Twitter user's neighborhood to infer attributes of that Twitter user herself. To our knowledge, this is an entirely unexplored direction in the latent attribute inference literature. Despite lack of investigation, the principle of homophily provides reason to believe it is a rich source of information about an individual. Homophily, often summarized using the moniker "birds of a feather flock together", is the tendency for individuals to seek out and associate with others who share similar attributes (e.g., beliefs, physical features, and activities) (McPherson, Smith-Lovin, and Cook). While this phenomenon is far from universal (consider the heterosexual bias in mate selection), it is believed to be a major mechanism of social organization in both physical and virtual settings (McPherson, Smith-Lovin, and Cook ; Thelwall 2009). Under the hypothesis that individuals with similar attributes cluster together in online social networks, we augment established user-centric latent attribute inference methods with information about user neighborhoods and assess the effect that the addition of this information has on the overall accuracy of attribute inference.

We assess the contribution of neighborhood feature data to the inference of three attributes: gender, political orientation, and age. These three attributes exhibit different degrees of assortativity, allowing us to evaluate how assortativity affects the performance of our method (Thelwall 2009; Conover et al. 2011b).

For both age and political orientation, we observe that augmenting features of the user with features of a subset of their neighborhood boosts inference accuracy by at least 3% — which, given the baseline performance of user-only feature vectors, constitutes a 20% to 35% improvement towards perfect inference. This performance boost in both attributes (both of them known to be assortative in the network) indicates that neighborhood data may stand to improve the inference of many different attributes.

It is noteworthy that different definitions of neighborhood maximally improve the inference accuracy for different attributes (closest friends for gender, all friends for politics, and least popular friends for age). This suggests that, while neighborhood context is useful, one must be attentive to the precise neighbors that are being used.

We also find that, if the user features are omitted entirely, features from the user's neighborhood *alone* produce inference accuracy that matches or surpasses those produced by user-only feature vectors. Only in the case of political assortativity do neighborhood features alone beat user features. These results indicate that, given the choice, neighborhood data can be comparable or better than user data at inferring a user's age and political orientation.

Related Work

Of existing latent attribute inference efforts, we are aware of only two which have touched on the idea of using neighborhood context to boost inference accuracy (Pennacchiotti and Popescu 2011; Conover et al. 2011a). In both, accounts already known to be associated with one of two attribute-labels (e.g., *Democrat* and *Republican*) were identified. In training and testing data, user feature vectors included fea-

tures indicating the degree of their connectivity to these different pre-labeled users. In (Conover et al. 2011a), this was done by identifying strong partisan clusters in the Twitter network prior to building user feature vectors. In (Pennacchiotti and Popescu 2011), commonly friended and mentioned accounts were identified among users with a given attribute (e.g., the accounts of party leaders among Republicans and Democrats).

Data and Methods

Data

Each attribute dataset consisted of approximately 400 labeled Twitter users, 200 with one label (e.g., "female") and 200 with a second label (e.g., "male"). In addition, all of the friends of these labeled users were identified as wellFor each of these labeled and neighbor users, the most recent 1000 tweets generated by the user were collected. This comprised a single dataset. It is worth noting that the scale of the datasets (400 users) is much smaller than those used in prior work. This is due in large part to the amount of neighborhood data that needed to be collected for each user — each dataset required more than 100GBs of storage in MySQL.

In order to collect the datasets, it was necessary to (1) decide on the two contrasting labels that would be used and (2) identify users that could be reliably assigned one label or the other. These decisions were specific to the attribute of interest, as described next. In each case, manual inspection of the accounts collected confirmed that the label assignments were correct.

Gender. In this case, the labels were self-evident: *male* and *female*. In order to identify male and female users, we used a technique proposed in (Mislove et al. 2011): we found Twitter accounts for which (1) the user had given a full (first and last) name and (2) the user's first name was one of the top 100 most common names on record with US social security department for baby boys/girls born in the year 2011. 192 male and 192 female labeled users were collected.

Age. We chose to distinguish between individuals aged 18 - 23 (hereafter "18+") and 25 - 30 ("25+"). In order to identify labeled individuals, we used the Twitter spritzer to collect all tweets in which an individual announced his or her own birthday (e.g., "Happy ##th/st/nd/rd birthday to me"). Retweets were ignored. Ultimately, 192 18+ users and 194 25+ users were labeled and collected.

Political orientation. Political users were identified as either Republican or Democrat. Following other prior work analyzing political discourse through Twitter, we identified users labeled as either Republican or Democrat on the `wefollow` website (`http://www.wefollow.com`) (Pennacchiotti and Popescu 2011). 200 Republican and 200 Democrat labeled users were collected.

Machine Learning Framework

A user's feature vector consisted of a set of N features computed over his microblog content and the same set of N features computed over the microblogs of a subset of his immediate friends. Some features required pre-processing a subset of labeled users (e.g., in the case of discovering the k-top most discriminating words used by one class of users). It is important to note that neighbor features were always *identical* to those used for the labeled users themselves (e.g., the k-top words evaluated for users were the same words evaluated for the neighbors).

All neighborhood features were the average value of that feature over each of the neighbors (the feature was computed for each neighbor and then averaged over all neighbors).

In order to thoroughly characterize the contributions of neighborhood data to attribute inference accuracy, we tested (1) using different subsets of a user's neighborhood (neighborhood policies) and (2) different ways of combining the user's own features and his neighborhood's features. Each combination of neighborhood policy and feature merging strategy (called a *configuration*) yielded a different dataset. For a given attribute, each configuration dataset had the same number of users, though the presence of neighborhood features and the way these were combined with the user-specific features for each labeled user differed. The neighborhood policies and user-neighbor combining strategies will be covered in a later subsection.

Initially we considered both support vector machines and gradient boosted decision trees (e.g., (Pennacchiotti and Popescu 2011; Burger, Henderson, and Zarrella 2011)) as binary classifiers for each attribute. SVMs consistently outperformed GBDTs, thus here we only report results from SVMs. We used the established SVM library, `libSVM` (Chang and Lin 2011). We use the radial basis function as the SVM kernel with cost and gamma parameters chosen using a grid search technique.

Features

Prior work in this area offers a rich set of different user features based on tweet text. We chose to include almost all of them as well as some new ones we devised.

k-top words. The k most differentiating words used by each labeled group were included as individual features (Pennacchiotti and Popescu 2011; Rustagi et al. 2009; Burger, Henderson, and Zarrella 2011).

k-top stems. Plurals and verb forms can weaken the signal obtained from k-top words by causing forms of the same word to be handled as separate words (e.g., "houses", "housing", and "house" are all derived from the stem "hous"). To address this, we passed all words through the Lovins stemmer and obtained the k-top differentiating stems for each labeled group (Lovins 1968).

k-top digrams and trigrams. In the training data, the k most differentiating digrams and trigrams were identified for both labels (Rao et al. 2010; Burger, Henderson, and Zarrella 2011).

k-top co-stems. In prior work, the ends of words (e.g., conjugations, plurals, and possessive marks) were shown to give notable signal about a variety of blog author attributes (Lipka and Stein 2011). These strings, called *co-stems*, can be obtained by subtracting the stem returned by the Lovins

stemmer and processing only the ending that remains (i.e., the word minus the stem).

k-top hashtags. Hashtags operate as topic labels. Prior work has shown that the extent to which topics are attribute-specific, they can be used for attribute inference (Pennacchiotti and Popescu 2011; Conover et al. 2011a).

Frequency statistics. We also included a number of frequency statistics (some of which appeared in (Rao et al. 2010)): tweets, mentions, hashtags, links, and retweets per day.

Retweeting tendency. The extent to which an individual propagates information was included by computing the ratio of retweets to tweets (Rao et al. 2010).

Neighborhood size. The ratio between number of followers and number of friends has been used as a measure of a user's tendency towards producing vs. consuming information on Twitter (Rao et al. 2010). We incorporated this as a feature as well.

Note that, while the neighborhood size, retweeting tendencies, and frequency statistics could be computed directly for an arbitrary user, the k-top words/n-grams/stems/co-stems/hashtags were attribute-specific and had to be pre-computed from a separate set of labeled users. This was done prior to computing any feature vectors for a given attribute. A value of $k = 20$ was used for all k-top features. Small changes to k did not affect overall performance, though small values of k introduced noise and much larger values of k made it difficult to obtain meaningfully differentiating terms.

Neighborhood Policies

In this study, we evaluated four different policies for selecting friends for inclusion in a user's neighborhood feature set. It is noteworthy that we included only friends (not followers) in this analysis since these correspond most directly to the principle of homophilic association (in Twitter, a user can exercise greater selectivity over who she follows than who follows her).

All. Under this policy, all of a Twitter user's friends were included.

n-most popular. Here, (hereafter, *Most*), the n-most popular friends were selected. Popularity was assessed in terms of number of followers.

n-least popular. Under this policy (*Least*), the n-least popular friends were selected. In this case, we are favoring the individuals in a user's neighborhood who have the fewest followers.

n-closest. Under this policy (*Closest*), the n-closest friends were selected. Closeness is not a directly observable property in twitter (or any other social platform, to our knowledge). The closest friends, therefore, were identified by determining the n friends whom the user mentioned most times in tweets (including retweets).

Configuration	Age	Gender	Political
UserOnly	0.751	0.795	0.890
Nbr-All	0.736	0.669	0.920
Nbr-Most	0.619	0.688	0.777
Nbr-Least	0.691	0.560	0.725
Nbr-Closest	0.716	0.598	0.895
Avg-All	0.795	0.750	0.918
Avg-Most	0.739	0.749	0.885
Avg-Least	0.805	0.758	0.878
Avg-Closest	0.779	0.674	0.909
Join-All	0.764	0.799	0.932
Join-Most	0.741	0.755	0.889
Join-Least	0.782	0.774	0.873
Join-Closest	0.772	0.802	0.915

Table 1: The overall accuracy of the SVM-based classifiers on datasets constructed using different combinations of user and neighborhood data. The top row, *UserOnly*, corresponds to results obtained from feature vectors that contained only data from the user's microblog content. All other rows involve configurations that incorporated neighborhood data.

A value of $n = 20$ was used throughout this study. As with k, we found that small changes to n made no difference. However, small values of n made the neighborhood subsample too small to be a reliable source of signal and large values of n exhausted the pool of candidate neighbors that could be included (leading to variability in the number of neighbors actually being included in the subsample).

Merging User and Neighborhood Features

Features were computed for users and subsets of their neighborhoods, per the different neighborhood policies described above. Since we wanted to understand the extent to which neighborhood data improved inference accuracy for each attribute, we designed four different kinds of feature vectors: *user-only* (all neighborhood features were omitted), *neighbor-only* (only the neighborhood features were used), *averaged* (each feature was the average over the neighborhood and the user feature values), and *joined* (the user features and neighborhood features were concatenated). In effect, user-only, neighbor-only, and averaged all had the same sized feature vectors; joined datasets had feature vectors that were twice the size of the other three (since user and neighborhood features were concatenated to one another). Hereafter, the combination of the kind of merged feature vector (i.e., user-only, neighbor-only, averaged, and joined) and the neighborhood policy (i.e., all, most, least, and closest) is referred to as a *configuration*.

Results

For each attribute, we generated a dataset for each possible configuration (enumerated in the left column of Table 1) from the Twitter data collected. A 10-fold cross-validation was done to assess the performance of the SVM-based classifiers at inferring the attribute of interest. The results are shown in Table 1. Standard deviations (not shown due to space limitations) are small enough to statistically support key differences.

Discussion

In this project, we evaluated the extent to which features present in a Twitter user's immediate neighbors can improve the inference of attributes possessed by the user herself. We considered this for three different attributes — age, gender, and political orientation — and found that substantial signal does exist. Our results support several noteworthy conclusions, which we discuss here.

Neighborhood context improves inference accuracy. Table 1 shows that adding neighborhood data improved the accuracy of inferred attribute labels for both age and political orientation (the results for gender will be discussed in detail below). These improvements were not only statistically significant, but also sizable. While absolute accuracy changes of 3% to 5% might seem modest, we contend that it is also important to consider the degree of improvement towards perfect inference, the ultimate goal. Under this perspective, our methods have improved the quality of inference by 21% (age) and 38% (political orientation).

Attribute assortativity influences accuracy gain. The inclusion of neighborhood features yielded improvements in inference accuracy for only two attributes. These improvements varied in degree: gender showed no statistically significant improvement, political orientation improved by 0.042, and age benefitted most (0.054).

Since assortativity quantifies the degree to which an individual is exclusively surrounded by like (or different) individuals, we should expect attributes with high observed assortativity (or disassortativity) to benefit from the inclusion of neighborhood features, and those with little or no observed assortativity to benefit much less. This is precisely what we observe. Gender has been reported to have minimal assortativity both in the physical and online social networks (McPherson, Smith-Lovin, and Cook ; Thelwall 2009). This explains the lack of improvement made our method on gender. Age and political affiliation have both been shown to be highly assortative (McPherson, Smith-Lovin, and Cook ; Conover et al. 2011b; Adamic and Glance 2005).

Choice of neighbors influences accuracy improvement. Table 1 reveals that the same neighborhood policy did not yield the best inference accuracy for all three attributes. In the case of age, the least popular neighbors yielded the best performance; for gender it was a tie between all neighbors and only the closest; for political orientation, the best performance was achieved by using all neighbors. This general finding underscores the fact that using neighborhood data to infer attributes requires attention to the selection of neighbors.

Neighborhood data can be comparable to user data. Comparing the *UserOnly* and *NbrOnly* accuracy values in Table 1, we find that when user-specific features are omitted, the neighborhood features are sufficient to obtain equally good or better inference accuracy for age and political orientation.

From a practical perspective, this means that our method provides a way of inferring attributes for protected users from their public neighbors. Provided that the attribute of interest has a moderate to high degree of assortativity, the features of the user's neighbors can provide an accurate indication of the user's own attribute value. Given that all existing methods heavily depend on the user's own microblog profile and content, this is a new capability with many practical applications in research and industry.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*.

Burger, J.; Henderson, J.; and Zarrella, G. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Conover, M.; Gonalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011a. Predicting the politial alignment of twitter users. In *Proceedings of the International Conference on Social Computing*.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Menczer, F.; and Flammini, A. 2011b. Political polarization on twitter. In *Proceedings of the International Conference on Weblogs and Social Media*.

Lipka, N., and Stein, B. 2011. Classifying with co-stems. In *Proceedings of the European Conference on Information Retrieval*.

Lovins, J. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11(1):21–31.

McPherson, M.; Smith-Lovin, L.; and Cook, J. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*.

Mislove, A.; Lehmann, S.; Ahn, Y.; Onnela, J.; and Rosenquist, J. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the International Conference on Weblogs and Social Media*.

Pennacchiotti, M., and Popescu, A. 2011. A machine learning approach to twitter user classification. In *Proceedings of the International Conference on Weblogs and Social Media*.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the International Workshop on Search and Mining User-generated Contents*.

Rustagi, M.; Prasath, R.; Goswami, S.; and Sarkar, S. 2009. Learning age and gender of blogger from stylistic variation. In *Proceedings of the International Conference on Pattern Recognition and Machine Learning*.

Thelwall, M. 2009. Homophily in myspace. *Journal of the American Society for Information Science and Technology*.